Which Type of Research is Cited More Often in Wikipedia? A Case Study of PubMed Research

Tahereh Dehdarirad*, Fereshteh Didegah** & Hajar Sotudeh***

* Department of Communication and Learning in Science, Chalmers University of Technology, Sweden

** iSchool, University of British Columbia / Scholarly Communications Lab, Simon Fraser University, Vancouver BC, Canada

*** Department of Knowledge and Information Sciences, Shiraz University, Iran

STI conference, Leiden, September 2018

Introduction

- Wikipedia is a prominent source of general healthcare information, extensively used by the general public, students, and health care professionals (Kousha & Telwall, 2016).

- More than 155,000 Wikipedia medical articles, written in different languages, were viewed more than 4.88 billion times in 2013, making it one of the most viewed medical and health care resources on the internet (Heilman & West, 2015).

- Given its popularity, it is important to ensure content quality of Wikipedia articles, which could be measured to an extent through articles' references.

Aim

This research aims to study the characteristics of external sources cited in Wikipedia articles, in order to determine the reasons why some documents are selected as reliable sources for Wikipedia and others are not.

Research Questions

- Which document types are cited more often in Wikipedia?
- Are open access documents cited more than non-open access documents in Wikipedia? Which types of open access documents are favored?
- Which Medical Subject Headings (Mesh) are cited more often in Wikipedia?
- Which F1000 classes are cited more often in Wikipedia?
- Are there significant correlations between Wiki citation counts and F1000 counts, news counts, and tweet counts?

Methodology

Data collection and processing

- The current study is based on a random sample of publications from PubMed proportionally gathered from 1996 to 2017, which accounted for 3,905,323 records.

- Using PMID, a search was made in *Altmetric.com* (October 2017 version) *for the Wikipedia citations* of the corresponding documents. From this, 384,394 (~10%) PMIDs were cited at least once in Wikipedia (*cited set*) , while the rest of PMIDs (3,520,929) were not cited.

- For comparison purposes, a random sample of uncited documents was also selected proportionally from 1996 to 2017, which accounted for 371,521 documents (*uncited set*).

Methodology

Data collection and processing

- All types of documents were taken into account for this study.
- Open access status of publications was obtained from *Unpaywall.org*.
- MESH subject headings for each record were obtained from PubMed. In this paper, we refer to these headings as topics.
- F1000, news and tweet counts were obtained via Altmetric.com for both collections of cited and uncited publications.
- To answer question 4, documents were classified into six F1000 classes: new finding, confirmation, technical advance, controversial, novel drug target and good for teaching.

Methodology

Statistical procedures

- To compare the percentage of cited OA documents in both cited and uncited sets in Wikipedia, a two-sample proportion test was used.

- Similarly, to compare the percentage of F1000 classes between cited and uncited sets, two-sample proportion tests were used.

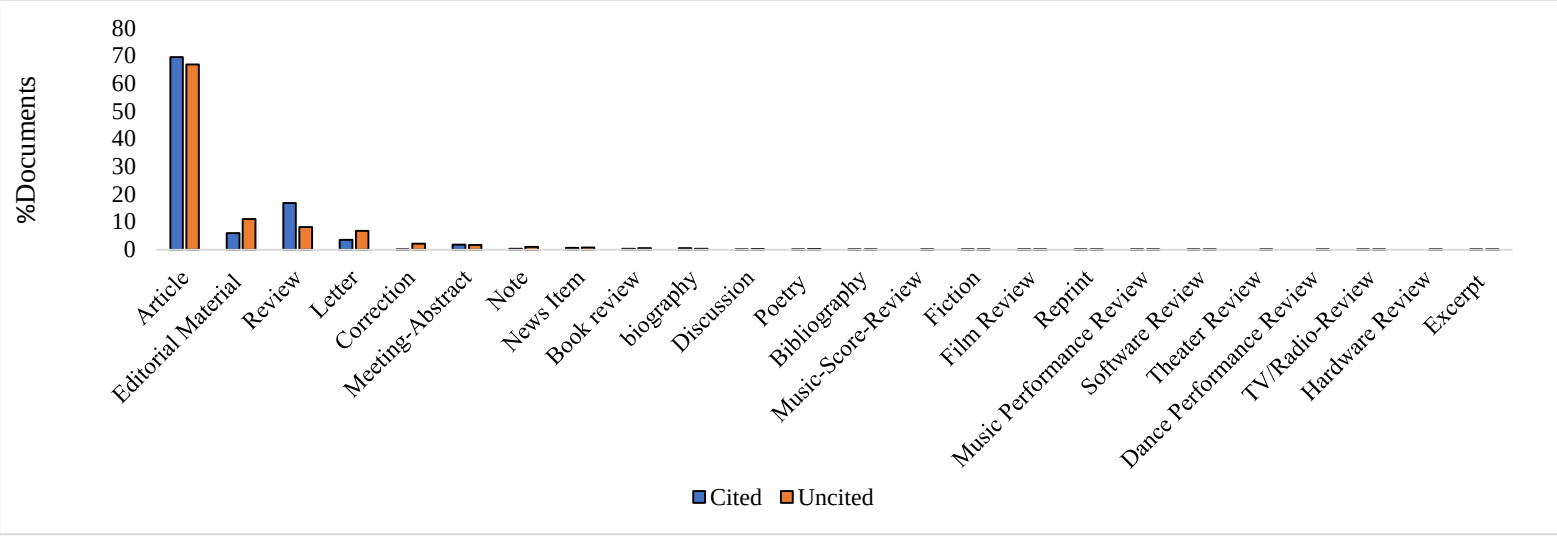
- Three Spearman correlations were used to study the relationship between Wikipedia citation counts, F1000 counts, news counts, and tweet counts for the entire collection of cited and uncited documents in Wikipedia.

Results

Question 1. Which document types are cited more often in Wikipedia?

- In both the cited and uncited sets of documents, editorial materials, reviews and letters are the top document types.
- The percentage of articles and reviews is slightly higher in the cited document set than the uncited set, though non-significant.

Figure 1: Document types for cited and uncited sets of documents in Wikipedia

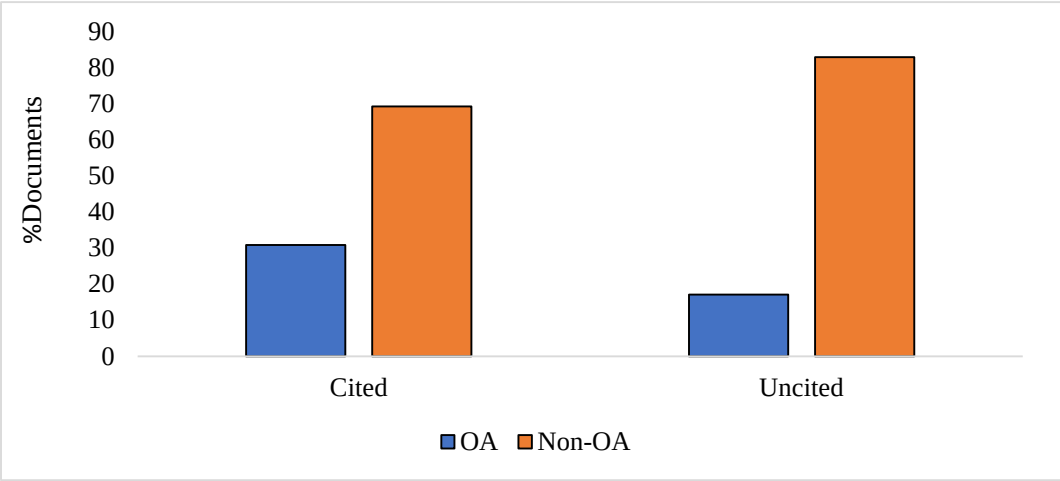


Results

Question 2. (a) Are open access documents cited more than non-open access documents in Wikipedia?

- Whilst around 30% of the cited set is open access, less than 20% of the uncited set is found to be open access. The percentage of cited OA documents is significantly higher than that of the uncited set[P<0.0001].
- More than 70% of documents in both sets are not open access.

Figure 2: OA status for cited and uncited sets of documents in Wikipedia.

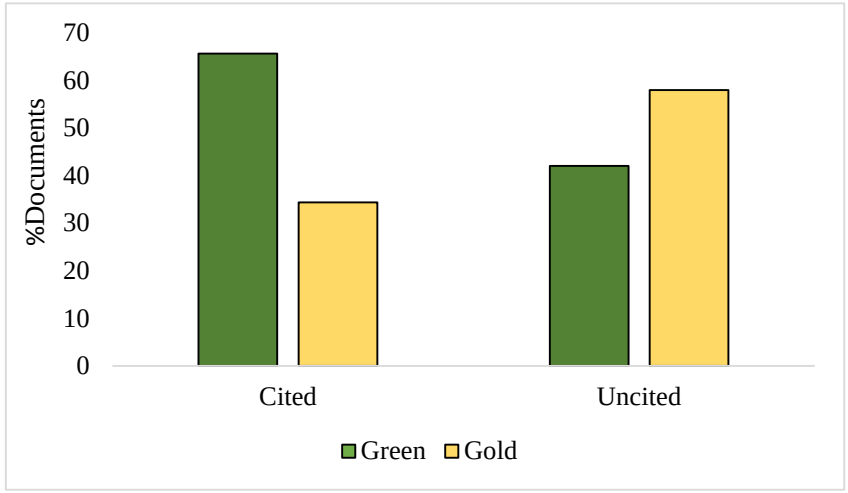


Results

Question 2.(b) Which types of open access documents are favored?

- For the cited set, more green type documents are found, whereas for the non-cited set, there are more gold type documents.

Figure 3: OA models for cited and uncited sets of documents in Wikipedia.



Results

Question 3. Which Medical Subject Headings (Mesh) are cited more often in Wikipedia?

- The cited set was classified into 15,852 topics, and the uncited set was classified into 10,289 topics.
- Neoplasms, Tuberculosis and Disease are the top three topics in both sets.

Table 1. Top 10 topics and their corresponding percentages for cited and uncited sets.

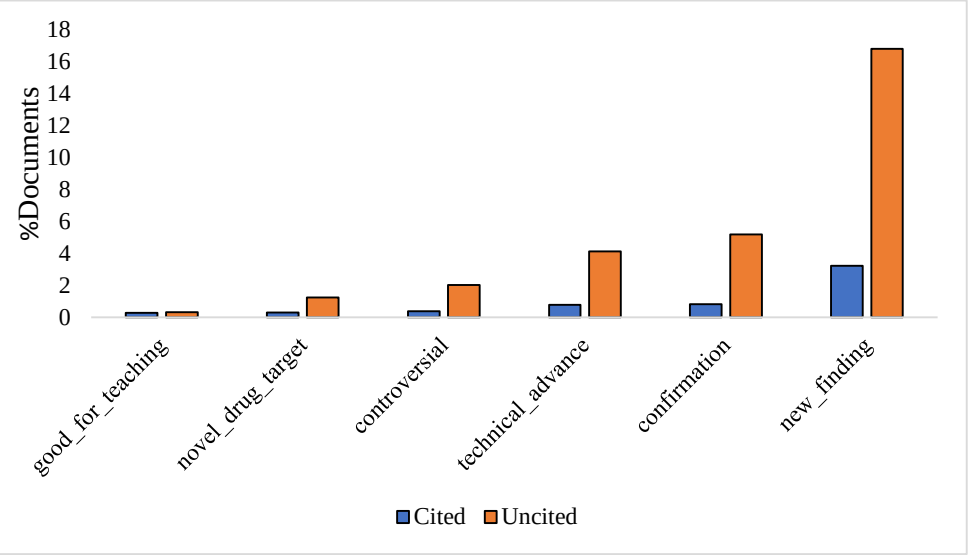
Cited	**Number (%)**	**Uncited**	**Number (%)**
Neoplasms	3631 (0.94)	Neoplasms	3503(0.94)
Tuberculosis	3139 (0.82)	Tuberculosis	3123 (0.84)
Disease	2737 (0.71)	Disease	2591 (0.70)
Mutation	2362 (0.61)	Medicine	1560 (0.42)
Biological Evolution	1958 (0.51)	Biometry	777 (0.21)
Phylogeny	1833 (0.48)	Intestines	768 (0.21)
Medicine	1651 (0.43)	Blood	766 (0.21)
Evolution. Molecular	1466 (0.38)	Brain	759 (0.20)
Gene Expression Regulation	1379 (0.36)	Anesthesia	733 (0.20)
Signal Transduction	1354 (0.35)	Tooth	730 (0.20)

Results

Question 4. Which F1000 classes are cited more often in Wikipedia?

- The majority of documents in both the cited and uncited sets of documents are classified into the '*new finding*' class of F1000.
- However, the proportion of uncited documents set in this class (~16%), is significantly higher than that of cited documents set (~4%; P<0.0001).

Figure 4: F1000 classes for cited and uncited sets.



Results

Question 5. Are there significant correlations between Wikipedia citation counts and F1000 counts, news counts, and tweet counts?

- Whilst a significant negative correlation is found between Wikipedia citation counts and F1000 and tweet counts, a very weak positive correlation is found between Wikipedia citation counts and news counts.
- Whilst 9.71% of documents cited in Wikipedia are mentioned in news outlets, only 7.13% of uncited documents are mentioned in news outlets.

Table 2. Spearman's rho correlation coefficients for the relationship between Wiki citation counts, F1000, news, and tweet counts.

Variable	F1000 post count	News post count	Tweet post count
Wiki citation count	-0.26*	0.07*	-0.35*

* p <0.0001

Conclusions

- A document type similarity for both the cited and uncited sets of documents, with the articles, reviews and editorial materials being more visible.
- Whilst the documents cover a broad range of topics, the top three topics are the same in both sets.
- The open access status of documents indicates that Wikipedia favors OA documents, although a large number of cited documents are non-OA.
- Regarding the F1000 classes, the majority of both the cited and uncited documents are categorized as “new finding”.
- Finally, our findings show significant, although weak correlations between Wiki citation counts and F1000, tweet and news counts. Whilst F1000 and tweet counts correlate negatively with Wikipedia citation counts, the news counts have a positive correlation.

Discussion

- Overall, the editors of English Wikipedia in medicine act as “distillers” of quality science.

- They interpret and distribute open/closed access knowledge to a broad public audience via different document types, whilst focusing on new findings and current medical knowledge.

- Moreover, it seems that Wikipedia’s focus is neither specialized, nor generalized, but it is something of a rather “general scientific” nature.

References

- Björk, B. C., & Paetau, P. (2012). Open access to the scientific journal literature–status and challenges for the information systems community. Bulletin of the Association for Information Science and Technology, 38(5), 39-44.
- Didegah, F. (2017). Factors associated with Wikipedia citations vs. traditional citations to research articles. WikiCite Conference. Vienna, Austria, 23 May.
- Evans, P., & Krauthammer, M. (2011). Exploring the Use of Social Media to Measure Journal Article Impact. *AMIA Annual Symposium Proceedings*, *2011*, 374-381.
- Hajjem, C., Harnad, S., & Gingras, Y. (2006). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. arXiv preprint cs/0606079.
- Haigh, C. A. (2011). Wikipedia as an evidence source for nursing and healthcare students. *Nurse Educ Today, 31*(2), 135-139.
- Heilman, J. M., & West, A. G. (2015). Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *J Med Internet Res*, *17*(3), e62.
- Koppen, L., Phillips, J., & Papageorgiou, R. (2015). Analysis of reference sources used in drug-related Wikipedia articles. *Journal of the Medical Library Association: JMLA, 103*(3), 140–144.
- Kousha, K., & Thelwall, M. (2017). Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology, 68*(3), 762-779.
- Laakso, M. (2014). Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed. Scientometrics, 99(2), 475–494. Retrieved from http://link.spri nger.com/article/10.1007%2Fs11192-013-1205-3.
- Laurent, M. R., & Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association: JAMIA*, *16*(4), 471-479.
- Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday; Volume 12, Number 8 - 6 August 2007*.
- Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. arXiv preprint arXiv:1203.4745.

References

- Shafee, T., Masukume, G., Kipersztok, L., Das, D., Haggstrom, M., & Heilman, J. (2017). Evolution of Wikipedia's medical content: past, present and future. *J Epidemiol Community Health*, *71*(11), 1122-1129.
- Sotudeh, H., Ghasempour, Z., & Yaghtin, M. (2015). The citation advantage of author-pays model: The case of Springer and Elsevier OA journals. Scientometrics, 104(2), 581-608.
- Sotudeh, H., & Estakhr, Z. (2018). Sustainability of open access citation advantage: the case of Elsevier's author-pays hybrid open access journals. Scientometrics, 1-14.
- Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology, 68*(9), 2116-2127.
- Thelwall, M. (2016). Does astronomy research become too dated for the public? Wikipedia citations to astronomy and astrophysics journal articles 1996–2014. *El Profesional de la Información, 25*(6), 893–900.
- Unpaywall (2018). Retrieved from http://unpaywall.org.
- Xia, J. F., Myers, R. L. & Wilhoite, S. K. (2011). Multiple open access availability and citation impact. Journal of Information Science, 37(1), 19-28.
- Wikipedia (2008). Wikipedia: verifiability. Wikimedia Foundation. Retrieved 3 April 2018, from http://en.wikipedia.org/wiki/Wikipedia:Verifiability.
- Wikipedia (2018). Wikipedia: Identifying reliable sources (medicine). Wikimedia Foundation. Retrieved 5 April 2018, from https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources_(medicine).

Thank you!